UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Chair of Practical Computer Science IV:
Dependable Systems Engineering

Master's Thesis

Vulnerabilities in Privacy-Preserving Record Linkage: The Threat of Dataset Extension Attacks

as part of the degree program Master of Science Business Informatics submitted by

Marcel Mildenberger
Matriculation number 1979905

on February 7, 2025.

Supervisor: Prof. Dr. Frederik Armknecht
PhD Student Jochen Schäfer

Abstract

The abstract should serve as an independent piece of information on your Thesis conveying a concise description of the main aspects and most important results. It should not be excessively long.

Write the abstract.

Contents

List of Figures

List of Tables

List of Algorithms

List of Code Snippets

Acronyms

1 Introduction

Data and record linkage is an important aspect of research and software projects, enabling the integration of data from different sources about the same entity to gain additional insights. This is particularly important in sectors such as healthcare or social sciences. In the United States, for example, the fragmented healthcare and public health ecosystem can benefit greatly from effective data linkage. The COVID-19 pandemic highlighted the critical importance of timely, accurate and efficient data linkage, as the lack of it led to problems in integrating disease and vaccination data. In response, organisations such as the Centers for Disease Control and Prevention and the Food and Drug Administration have launched projects to address these challenges and further develop linkage techniques. [**pathak2024privacy**]

In scenarios such as the COVID-19 pandemic, data integration efforts often involve linking records related to natural persons across multiple sources. For example, integrating data from various healthcare providers, laboratories, and public health agencies typically requires the use of pseudo-identifiers derived from Personally Identifiable Information such as names, dates of birth, or other sensitive information. However, the reliance on Personally Identifiable Information introduces significant privacy concerns, as improper handling of such data can lead to the re-identification of individuals, potentially resulting in severe consequences such as data breaches, identity theft, or unauthorized access to personal health information.

To address these privacy risks, various techniques have been developed to protect Personally Identifiable Information, primarily by encrypting or encoding the data prior to linkage. Privacy-Preserving Record Linkage techniques are designed to facilitate data integration without exposing sensitive information, ensuring that datasets can be linked securely across different entities. However, the use of encrypted Personally Identifiable Information as pseudo-identifiers presents additional challenges. Key questions arise regarding how to efficiently encode sensitive information while preserving the ability to match records accurately. Furthermore, performing linkage on encrypted or encoded data requires specialized algorithms that can operate effectively without compromising the privacy of the underlying information [**<empty citation>**].

In order to still be able to perform linkage while preserving privacy, similarity preserving encodings are applied to the identifiers. Without such similiraty preserving encoding, matches between encrypted entities in different databases would not be possible. Over time, three main privacy-preserving encoding schemes have emerged as enablers for Privacy-Preserving Record Linkage. [**vidanage2020graph**; SAH24]

Bloom Filter encoding is the most widely used technique in Privacy-Preserving Record Linkage and is often regarded as the reference standard. Bloom Filters were introduced to Privacy-Preserving Record Linkage due to their simplicity and efficiency in both storage and the computation of private set similarities. Their compact representation and probabilistic nature make them ideal for scalable Privacy-Preserving Record Linkage systems. Further research has sought to enhance the security of Bloom Filter encoding by incorporating diffusion layers, which help obscure patterns and reduce vulnerability to certain attacks. Tabulation MinHash Encoding is a more recent encoding method that offers distinct advantages in specific use cases. While it is less commonly used in Privacy-Preserving Record Linkage, its

simplicity and efficiency in calculating set similarities make it a viable option. Compared to Bloom Filter encoding, Tabulation MinHash Encoding generally provides stronger security guarantees. However, this added security comes at the cost of increased computational complexity and memory usage. Two-Step Hash Encoding introduces a novel approach designed to address some of the limitations associated with Bloom Filter and Tabulation MinHash Encoding. This method employs a two-step process: initially encoding data into multiple Bloom Filters, followed by an additional hashing step that converts the encoded data into a set of integers for similarity comparisons. This layered approach enhances security while maintaining efficient similarity computations. [**vidanage2020graph**; SAH24]

In practice, Bloom Filter based Privacy-Preserving Record Linkage has become the dominant standard and is widely adopted in areas such as crime detection, fraud prevention, and national security. However, Privacy-Preserving Record Linkage systems are not without limitations and vulnerabilities. Previous research has demonstrated that several attacks exist targeting Privacy-Preserving Record Linkage systems, with a primary focus on exploiting the weaknesses inherent in Bloom Filter encodings. These attacks specifically target vulnerabilities in Bloom Filter constructions, such as the weaknesses introduced by double hashing, structural flaws in the filter design, and susceptibility to frequent pattern-mining techniques. Additionally, language model-based attacks and graph-based dictionary attacks have been employed to compromise Bloom Filter encoded data. Notably, no specialized attacks have yet been developed for Tabulation MinHash Encoding or Two-Step Hash Encoding encodings, which suggests that research has primarily concentrated on the more widely used Bloom Filter scheme. [**vidanage2020graph**]

Nevertheless, a more recent and sophisticated attack has emerged that exploits vulnerabilities common to all Privacy-Preserving Record Linkage encoding schemes. The Graph Matching Attack leverages publicly available data, such as phone books, to re-identify encrypted individuals based on overlapping records between plaintext and encrypted databases [**vidanage2020graph**; SAH24]. Unlike earlier attacks that focused solely on the encoding scheme of Bloom Filters, the Graph Matching Attack operates independently of the encoding scheme by solving a graph isomorphism problem to match records. While Graph Matching Attacks can successfully re-identify individuals present in both the plaintext and encrypted datasets, their effectiveness is limited to the overlapping subset of the two databases [SAH24].

This work aims to advance beyond traditional Graph Matching Attacks by re-identifying not only individuals present in the overlapping datasets but as many individuals as possible from the encrypted Privacy-Preserving Record Linkage data. To accomplish this, the newly introduced Dataset Extension Attack builds upon the foundation established by Graph Matching Attacks. The Dataset Extension Attack leverages a neural network trained on the subset of previously re-identified individuals to predict and decode the remaining encrypted records. By doing so, the Dataset Extension Attack significantly expands the scope and effectiveness of the attack, enabling broader de-anonymization of Privacy-Preserving Record Linkage datasets beyond the limitations of existing graph-based methods.

1.1 Motivation

The increasing use of Privacy-Preserving Record Linkage in highly sensitive domains such as healthcare, finance, national security, and fraud detection necessitates rigorous research to validate existing techniques and ensure robust data protection. As data-driven technologies

continue to evolve requiring approaches like Privacy-Preserving Record Linkage, so do the methods used by malicious actors to exploit these systems. While data privacy has always been a critical concern, its importance has surged in an era dominated by Artifical Intelligence and Machine Learning. These technologies increasingly rely on large datasets, many of which contain Personally Identifiable Information that, if compromised, could lead to significant privacy violations. Furthermore, the rise of data brokerage, where personal data is collected, aggregated, and sold—often without explicit user consent has intensified concerns around data misuse. As AI models grow more sophisticated, the demand for extensive, high-quality data escalates, making privacy protection a pressing issue [**ldc2024**; **cacgroup2024**; **arxiv2024**].

In this context, the vulnerabilities of Privacy-Preserving Record Linkage systems to emerging attack vectors become particularly alarming. As highlighted in the introduction, researchers have demonstrated that Privacy-Preserving Record Linkage systems are susceptible to Graph Matching Attacks, which exploit the similarity preserving properties of common encoding schemes to re-identify individuals. These attacks directly undermine the primary objective of Privacy-Preserving Record Linkage: to protect sensitive data during the linkage process. Although current Graph Matching Attacks are limited to re-identifying individuals present in both encrypted and plaintext datasets, this still represents a significant privacy risk, especially in domains where even partial data exposure can have serious implications.

The potential implementation of a Dataset Extension Attack introduces even greater risks. Unlike Graph Matching Attacks, which are confined to the intersection of datasets, Dataset Extension Attacks aim to extend the attack to re-identify as many individuals as possible from the encrypted database. By leveraging neural networks trained on previously decoded data, Dataset Extension Attacks can predict and decode additional records, potentially leading to the complete de-anonymization of entire encrypted datasets. This not only nullifies the privacy guarantees offered by Privacy-Preserving Record Linkage systems but also raises critical questions about the future viability of widely used techniques, such as Bloom Filter based Privacy-Preserving Record Linkage.

The primary motivation for this research is to proactively investigate and demonstrate the consequences of such advanced attacks to prevent their realization in real-world scenarios. By exposing the potential vulnerabilities of Privacy-Preserving Record Linkage systems, this work aims to highlight how attackers could exploit encrypted data to compromise privacy at scale. A successful implementation of the Dataset Extension Attack will reveal that state-of-the-art methods are insufficiently secure, emphasizing the urgent need for more robust privacy-preserving techniques. Moreover, the lack of comprehensive research into extending the scope of attacks beyond the intersection of datasets underscores the necessity of this study. While significant efforts have been made to address the vulnerabilities exposed by Graph Matching Attacks, there is a notable gap in understanding how these systems can be compromised on a broader scale. By addressing this gap, this thesis seeks to contribute to the body of knowledge on Privacy-Preserving Record Linkage vulnerabilities and serve as a foundation for future research aimed at fortifying these systems against increasingly sophisticated threats.

1.2 Related Work

The work of Vidanage et al. [**vidanage2020graph**] introduces a novel attack against Privacy-Preserving Record Linkage, known as the Graph Matching Attack. There the authors first provide an overview about Privacy-Preserving Record Linkage, encoding methods and present

attacks on systems so far. Their novel attack exploits the similarity preserving properties of commonly used encoding schemes, such as Bloom Filter, making it universally applicable across various Privacy-Preserving Record Linkage methods. By utilizing a graph-based approach and solving a graph-isomorphism problem, the Graph Matching Attack aligns nodes in similarity graphs to successfully re-identify individuals using a plain database [**vidanage2020graph**]. This work is critical for this thesis, as the Dataset Extension Attack builds upon the re-identified individuals produced by the Graph Matching Attack and is extending the attack.

Another key contribution comes from Schäfer et al. [SAH24], who revisit and extend the work of Vidanage et al. [**vidanage2020graph**] The authors provide a thorough reproduction and replication of the proposed Graph Matching Attack, uncovering an undocumented preprocessing step in the original codebase that unintentionally impacted the attack's success rate. This undocumented prepropcessing step was intended to improve performance but instead introduced implementation errors. Schäfer et al. addressed this issue by correcting the preprocessing and enhancing the Graph Matching Attack to improve both robustness and efficiency. Their improved implementation achieved higher re-identification rates compared to the original approach by Vidanage et al. [SAH24]. The work of Schäfer et al. is particularly relevant to this thesis, as their enhanced Graph Matching Attack implementation and corresponding code base serves as the foundation for the Dataset Extension Attack proposed in this thesis. Therefore small adjustments to the Graph Matching Attack implementation will be done to build upon it.

Currently there is no approach like the proposed Dataset Extension Attack in research present. Therefore this thesis aims to fill this gap.

1.3 Contribution

The contribution of this thesis is structured into three primary components. First, a comprehensive analysis of Privacy-Preserving Record Linkage systems is conducted, with particular emphasis on the three major encoding schemes: Bloom Filter Encoding, Two-Step Hash Encoding Encoding, and Tabulation MinHash Encoding Encoding. This analysis aims to highlight the fundamental principles, strengths, and vulnerabilities of each encoding method, setting the stage for the subsequent investigation of their susceptibility to privacy attacks.

Next, the current state-of-the-art Graph Matching Attack is analyzed, and its limitations are discussed in detail. Although Graph Matching Attacks have proven effective in re-identifying individuals within overlapping datasets, their applicability is restricted to the intersection of plaintext and encrypted records. This inherent limitation underscores the need for more advanced attack strategies that can extend beyond this.

The primary focus of this thesis is the implementation and evaluation of the Dataset Extension Attack, which seeks to surpass the capabilities of Graph Matching Attacks by decoding a larger proportion of encrypted records. To achieve this, the thesis delves into the conceptual foundations, theoretical underpinnings, and technical requirements of the Dataset Extension Attack. Building upon the initial re-identifications obtained through the Graph Matching Attack, the Dataset Extension Attack employs a supervised machine learning approach, specifically utilizing neural networks trained on previously decoded data to predict and re-identify remaining encrypted records. This method significantly broadens the scope of de-anonymization in Privacy-Preserving Record Linkage systems and provides a novel approach to the current research.

The Dataset Extension Attack is then evaluated against the three major Privacy-Preserving Record Linkage encoding schemes. While the specific encoding method has minimal impact on the Graph Matching Attack, which relies primarily on solving graph isomorphism problems, it plays a pivotal role in the Dataset Extension Attack. This is due to the fact that the neural network must be trained separately for each encoding scheme to account for the unique structural characteristics and encoding nuances. However, the Dataset Extension Attack is designed with adaptability in mind, ensuring that it can be effectively applied across different encoding schemes, thus enhancing its generalizability and practical relevance.

Through this research, the thesis aims to answer critical questions concerning the robustness of Privacy-Preserving Record Linkage systems. It investigates how effective supervised machine learning-based Dataset Extension Attacks are in re-identifying the remaining entries that Graph Matching Attacks cannot uncover. Furthermore, it explores how different encoding schemes influence the performance and accuracy of the Dataset Extension Attack, providing insights into which schemes are more susceptible to such attacks and why. By addressing these questions, the thesis contributes to a deeper understanding of the vulnerabilities inherent in Privacy-Preserving Record Linkage systems and lays the groundwork for developing more secure privacy-preserving techniques.

1.4 Organization of this Thesis

This thesis is divided into X main sections. First, an overview of Privacy-Preserving Record Linkage systems will be provided, with a particular focus on a thorough analysis of the most commonly used encoding techniques. Following this, the existing Graph Matching Attack will be introduced and explained to establish the foundation for the study. Additionally, an overview of neural networks will be presented to provide necessary background knowledge.

Next, a detailed description of the attack model for the Dataset Extension Attack will be outlined, including how neural networks are leveraged to enhance the attack. This is followed by an explanation of the actual implementation of the Dataset Extension Attack, along with a discussion of the experiments conducted. The results of the Dataset Extension Attack across different encoding schemes will then be analyzed and evaluated. Finally, the thesis will conclude with a summary of the key contributions, a discussion of the broader implications, and suggestions for future research.

2 Background

2.1 Overview of PPRL

2.2 Key encoding techniques

2.2.1 Bloom Filters

2.2.2 Tabulation MinHash

2.2.3 Two-Step Hashing

2.3 Graph Matching Attacks

3 Methodology

3.1 Conceptual framework for Dataset Extension

3.2 Implementation

4 Results

4.1 Analysis

4.2 Discussion

5 Conclusion

5.1 Summary

5.2 Future Work

Bibliography

[PSZ+24] Aditi Pathak, Laina Serrer, Daniela Zapata, Raymond King, Lisa B. Mirel, Thomas Sukalac, Arunkumar Srinivasan, Patrick Baier, Meera Bhalla, Corinne David-Ferdon, Steven Luxenberg, and Adi V. Gundlapalli. "Privacy Preserving Record Linkage for Public Health Action: Opportunities and Challenges." In: *Journal of the American Medical Informatics Association* 31.11 (2024). Advance access publication July 24, 2024, pp. 2605–2612. DOI: 10.1093/jamia/ocae196.

[SAH24] Jochen Schäfer, Frederik Armknecht, and Youzhe Heng. "R+R: Revisiting Graph Matching Attacks on Privacy-Preserving Record Linkage." In: *Proceedings of [Conference Name, if available]*. Available at: https://github.com/SchaeferJ/graphMatching. University of Mannheim. 2024.

A Auxiliary Information

Eidesstattliche Erklärung

Hiermit versichere ich, dass diese Abschlussarbeit von mir persönlich verfasst ist und dass ich keinerlei fremde Hilfe in Anspruch genommen habe. Ebenso versichere ich, dass diese Arbeit oder Teile daraus weder von mir selbst noch von anderen als Leistungsnachweise andernorts eingereicht wurden. Wörtliche oder sinngemäße Übernahmen aus anderen Schriften und Veröffentlichungen in gedruckter oder elektronischer Form sind gekennzeichnet. Sämtliche Sekundärliteratur und sonstige Quellen sind nachgewiesen und in der Bibliographie aufgeführt. Das Gleiche gilt für graphische Darstellungen und Bilder sowie für alle Internet-Quellen.
Ich bin ferner damit einverstanden, dass meine Arbeit zum Zwecke eines Plagiatsabgleichs in elektronischer Form anonymisiert versendet und gespeichert werden kann.

Datum	Marcel Mildenberger